Review of Economic Studies (2002) **69**, 589–610 0034-6527/02/00230589$02.00
© 2002 The Review of Economic Studies Limited

Fault Tolerant Implementation

KFIR ELIAZ
New York University

First version received 20 *March* 2000; *final version accepted* 28 *November* 2001 (*Eds.*)

In this paper we investigate the implementation problem arising when some of the players are "faulty" in the sense that they fail to act optimally. The planner and the non-faulty players only know that there can be at most k faulty players in the population. However, they know neither the identity of the faulty players, their exact number nor how faulty players behave. We define a solution concept which requires a player to optimally respond to the non-faulty players regardless of the identity and actions of the faulty players. We introduce a notion of fault tolerant implementation, which unlike standard notions of full implementation, also requires robustness to deviations from the equilibrium. The main result of this paper establishes that under symmetric information any choice rule that satisfies two properties—k-monotonicity and no veto power—can be implemented by a strategic game form if there are at least three players and the number of faulty players is less than $\frac{1}{2}n - 1$. As an application of our result we present examples of simple mechanisms that implement the constrained Walrasian function and a choice rule for the efficient allocation of an indivisible good.

1. INTRODUCTION

Implementation theory studies the problem of a planner who faces a set of agents and wishes to associate a set of outcomes with each possible profile of the agents' preferences (the correspondence that assigns a set of outcomes to each profile of the agents' preferences is called a *choice rule*). The standard approach implicitly assumes that each agent is able to correctly choose his most preferred action. A question arises as to how robust are the conclusions reached by standard models to slight deviations from the full rationality assumption. If we believe that decision makers might err, we may be interested in constructing mechanisms that are immune to possible mistakes that some of the players might make.

This paper explores the question of implementation that is robust to the potential of having a limited number of agents who make mistakes. An agent who makes mistakes is viewed as a decision maker who has well defined preferences that are known to others, but who fails for one reason or another to behave optimally. We refer to such an agent as being *faulty*. For an observer who knows the preferences of a faulty player, it would seem as if the player was not acting in accordance to his preferences.

We concentrate on two aspects of error-prone behavior. First, incorrect decisions are hard to predict. There are also many potentially incorrect decisions that may be made. Second, the tendency to make mistakes is usually an unobservable feature of a decision maker. We may have a very good idea of what the preferences are of an individual (more money is preferred to less, less pain is preferred to more), but we may have no clue as to how successful this individual is in making the correct decisions.

The presence of faulty players introduces several complications to the problem of implementation. If players are aware of the possibility that some players might err, they may take that into account when considering their course of action. For example, players may wish to exploit the potential faultiness of some of the players. Thus, the strategic reasoning of the players may be affected by the mere knowledge that some players may be faulty. In addition, different players may entertain different beliefs over the identity and behavior of the faulty players. This raises the

question of what is an appropriate notion of equilibrium in such a setting. An equilibrium may at best describe some stable pattern of behavior by players who are potentially non-faulty. The planner must take into account that ultimately, the players who turn out to be faulty will behave in an unpredictable manner and may choose an action contrary to their incentives.

Our objective is to enable a planner to implement choice rules even if some of the agents he faces are faulty players. We consider the simple case in which the planner is restricted to using only strategic game forms, and where the agents know their own and other agents' true profile of preferences but the planner does not. We view the preferences of an individual and his tendency to make mistakes as two independent characteristics in the sense, that one cannot infer an agent's tendency to make mistakes from observing his preferences. We therefore assume that the planner and the non-faulty players only know that there can be at most k faulty players in the population. However, they know neither the identity of faulty players, their exact number nor how faulty players behave. Assuming that the non-faulty players' behavior conforms with an appropriate solution concept, the planner's objective is to construct a game form such that for every possible profile of the agents' preferences and regardless of the actions of the faulty players the following is satisfied. The set of outcomes, in which any subset with at least $n - k$ players execute their equilibrium strategies, equals the set of outcomes that the choice rule associates with the profile of preferences. A game form having this property is said to achieve *fault tolerant implementation* of the choice rule.

It is important to note that faulty players do not necessarily represent players who make mistakes. The two central features of faulty players are the unpredictable nature of their behavior and their unknown identity. Thus, faulty players may also be interpreted as "malicious" agents who wish to prevent the planner from attaining his goal. The malicious agents cannot be identified by the planner or by the normal agents as their preference for harming the mechanism is unobservable to others. Another interpretation of faulty players is that of a minority with unknown preferences. This interpretation considers the non-faulty players to be the "majority", a homogeneous group of agents with identical preferences, who only know the preferences of their own type. From their point of view, any agent who does not share their preferences may have any preferences.

Fault tolerant implementation can be interpreted as a criterion of robustness against the worst possible mistakes, mistakes, which would cause some player to deviate from truth-telling. An alternative interpretation is that the planner requires robustness to *any* possible beliefs that players might have about the behavior of faulty players. By imposing such a strong robustness requirement, we do not need to make explicit assumptions on the set of possible mistakes and on the probabilities of making mistakes and of being faulty.

Fault tolerant implementation can be relevant in several contexts. Some examples include the following.

(1) *Costly analysis of possible scenarios*—A fault tolerant mechanism may be preferred in situations which require robustness to mistakes where it is costly for the planner to identify all the reasonable mistakes and all the reasonable beliefs that players may have about the faultiness of others.
(2) *Robustness to players with preferences over a richer domain*—There may be situations in which the preferences of some of the players may be defined over a domain, which is richer than the set of alternatives offered by the mechanism. For example, in the recent Israeli elections there were only two possible outcomes: Ehud Barak or Ariel Sharon. Since Barak represented the pro-peace movement led by the left wing, the Israeli Arabs clearly preferred Barak to Sharon. Therefore, the decision of the Israeli Arabs not to vote (which in effect is a decision to vote for Sharon) was clearly inconsistent with this group's

preferences over the set of candidates. However, the Israeli Arabs preferred an outcome, which was not offered by the voting mechanism, to their most preferred election outcome: they preferred to display their ability to punish a government official who mistreated them (Barak) even at the cost of having their least preferred candidate win the elections. Thus, fault tolerant implementation offers robustness against the possibility that the behavior of some players may not be rationalized by their ranking of the alternatives offered by the mechanism.

(3) *Discriminatory mechanisms*—There are situations in which a planner cares about the preferences of only a subset of all the agents in the sense that his choice rule is defined on the preferences of only the agents in this subset. The agents in this "privileged" subset share some common unobservable characteristics such that each member of this subset cannot be identified by the planner or by the other privileged agents. For example, these characteristics might be the religion of the agents, their political affiliation or their preferences on objects, other than the ones selected by the mechanism. In most cases the planner may not be able to condition participation in the mechanism on having certain characteristics (these characteristics may not be verifiable). Under certain conditions (see the discussion in the final section of this paper), fault tolerant implementation allows a planner to design a mechanism, which is robust to the behavior of the non-privileged agents.

The study of fault tolerant implementation requires a new notion of equilibrium and a new notion of implementation. Most equilibrium notions for strategic game forms require a player to optimally respond to the equilibrium strategies of the other players. However, if a player knows that some of the other players might not follow their equilibrium strategies, he may wish to deviate from his own equilibrium strategy. The presence of faulty players requires a different notion of strategic stability, one that takes into account possible deviations of players from their associated equilibrium strategies. We therefore define the notion of a *k-fault tolerant Nash equilibrium* (*k*-FTNE), which is viewed as a suggested course of action for each player in the game and which is stable in the following sense. Each non-faulty player has no incentive to deviate from the action suggested to him, regardless of the identity and actions of the faulty players, as long as $n - k - 1$ non-faulty players adhere to the actions suggested to them.

The notion of *k*-FTNE captures the following considerations. A *k*-FTNE strategy is a best response of a player for *any* belief over the faulty players. For any realization of faulty players in the population, any player who turns out to be non-faulty has no incentive to deviate from his associated *k*-FTNE strategy. Thus, the notion of *k*-FTNE is independent of the beliefs that players might hold regarding the faulty players, it does not require any assumptions regarding the nature of faulty behavior and it applies to any distribution of faultiness in the population.

The standard notion of implementability assumes that each player chooses his associated equilibrium strategy. However, even if we assume that every non-faulty player chooses his associated equilibrium action, a faulty player does not act according to the equilibrium. Thus, the set of possible outcomes in a game with at most *k* faulty players is not equal to the set of *k*-FTNE outcomes. It also contains the set of outcomes corresponding to action profiles in which the faulty players deviate from their equilibrium strategies. Therefore, the standard definition of full implementation does not guarantee that the only possible outcomes in a mechanism are the "desirable" ones.

We introduce a notion of implementability that is robust to the deviations of faulty players from the *k*-FTNE action profile. A game form implements a choice rule in *k*-FTNE if it satisfies two requirements. The first requirement is the standard definition of full implementation, which applies to the case in which none of the players are faulty. For every profile of preferences the

set of all k-FTNE outcomes must equal the set of outcomes selected by the choice rule. The second requirement calls for robustness to deviations from an equilibrium. For every profile of preferences and for every k-FTNE the set of outcomes selected by the choice rule must include the set of outcomes resulting from action profiles that differ from the k-FTNE by at most k actions.

Equipped with this framework for fault tolerant implementation, we can characterize the choice rules that can be implemented in k-FTNE. The necessary and sufficient conditions for k-FTNE implementation are closely related to the conditions found by Maskin (1999) to be necessary and sufficient for Nash implementation. We show that any choice rule that is *k-monotonic* and satisfies *no-veto-power* can be implemented in k-FTNE as long as the following conditions are met: (1) there are at least three agents participating in the mechanism and (2) a non-faulty player cannot form a majority by joining the set of all faulty players ($k < \frac{1}{2}n-1$). Conversely, any choice rule that is implementable in k-FTNE must satisfy *weak k-monotonicity*. The two notions of k-monotonicity are equivalent when applied to choice functions. Thus, k-monotonicity is both necessary and sufficient for k-FTNE implementation of choice functions that satisfy no-veto-power. With regards to well-known choice rules, the constrained Walrasian correspondence is $(n-1)$-monotonic, whereas the Core is at least 1-monotonic.

The paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the k-FTNE solution concept and the notion of implementation in k-FTNE. In Section 4 we present the main results concerning the sufficient and necessary conditions required for k-FTNE implementation. Section 5 is devoted to applications of the concept of k-FTNE implementation in specific settings. We present a simple mechanism for implementing the constrained Walrasian function in a simple exchange economy setting. We also provide an example of a specific setting, in which a simple k-FTNE mechanism implements the efficient allocation of an indivisible good among n agents. We close with concluding remarks.

2. RELATED LITERATURE

Our work draws from the computer science literature on fault tolerant computation. This strand of literature investigates the design of protocols that allow parallel processing networks to continue carrying out correct computation, even if some of the processors are faulty. Faulty processors are either assumed to act in a totally unpredictable manner (for example, they can transmit false information), or they are assumed to be of limited computational complexity. There are two common interpretations of faulty processors of the first type (the type related to this paper): (a) parts of a network that do not function properly because of some technical problem or mistakes on part of the operators, and (b) components in the mechanism that are operated (perhaps in cooperation) by "malicious" agents who wish to bring down the network.

Most of the computer science literature focuses on increasing the efficiency of fault tolerant protocols and on raising the upper bound on the number of faulty processors in a network, above which the protocols cease to be fault tolerant. As long as the upper bound is met, the protocols are guaranteed to work with certainty, regardless of the exact number and identity of the faulty processors in the network (see Linial, 1994 for a survey of the main results in the theory of fault tolerant computation and their relevance to economics and game theory).

Note the main difference between the computer science approach to the design of protocols and the economic theory approach to the design of mechanisms: non-faulty processors are programmed according to a protocol and are assumed to follow the programmer's instructions; however, individuals participating in a mechanism cannot be programmed. To ensure compliance, a social planner must provide the agents with sufficient incentives to induce them to follow the prescribed course of action.

This paper attempts to respond to some of the criticism raised with respect to the reliance of classical implementation theory (see Moore, 1992) on substantive unbounded rationality of the agents. Thus, this paper represents a step towards incorporating a model of bounded rationality into implementation theory. Only a few works have addressed this issue. Sjostrom (1993) considers trembling-hand perfect implementation, which offers a different approach to modeling mistakes in implementation theory. Hurwicz (1986) studies Nash implementation when the agents' preferences are intransitive, cyclic or incomplete. Bartholdi *et al*. (1989) show that a certain class of voting schemes require excessive computation to determine the winner. An analysis of the implementation problem, which incorporates the theory of learning in games, has been carried out by Cabrales (1999), Cabrales and Ponti (2000) and Hon-snir *et al*. (1998). The idea that individuals may fail to respond in a contractual environment has been discussed in Segal (1999). He considers a principal-agent framework in the context of takeovers. He shows that if agents fail to respond to the principal's offer with a small probability, they become asymptotically non-pivotal and inefficiency obtains.

3. PRELIMINARIES

The triple $\langle N, C, P\rangle$ represents the *environment* in which the planner operates. Let N be a set of players $\{1, \ldots, n\}$, C a set of consequences, and P a set of preference profiles over C. An element of P will be denoted by p, a profile of preferences of a subset of players $M \subseteq N$ will be denoted by p_M, and a preference relation of a single player $i \in N$ by p_i. We use the notation $c \succsim_i b$ ($c \succ_i b$) to indicate that agent i weakly (strictly) prefers the outcome c to the outcome b. A *choice rule* that assigns a subset of C to each profile in P will be denoted by f, such that $f : P \to 2^C \backslash \{\phi\}$.

The planner controls the rules of the game, formalized as a game form. A *strategic game form with consequences in C* is a triple $\langle N, (A_i), g\rangle$, where A_i is the set of actions available to player i, and $a_i \in A_i$ is a single action of this player. We let $A = A_1 \times \cdots \times A_n$ with $a \in A$ denoting a profile of actions for the n players. The third component of the game form, $g : A \to C$, is an *outcome function* that associates an outcome with every action profile.

Faulty players

A player who does not act according to incentives will be called *faulty*. That is, given the actions chosen by the other players, a faulty player does not choose an action that leads to his most preferred outcome.

In the standard implementation framework, given a game form, an instance is a profile of the players' preferences. The players play the game induced by the given game form and the profile of preferences. We define an instance as a pair (p, k) where k is an upper bound on the number of faulty players in N. Given a profile of preferences, any subset of at most k players might play in an unpredictable manner. A non-faulty player knows that he is not faulty. However, he cannot tell whether another player is faulty or not, and he does not know the exact number of faulty players in N. He only knows that there cannot be more than k faulty players in N. The instance (p, k) is assumed to be common knowledge among the non-faulty players.

Given an instance (p, k), the planner only knows that there can be at most k faulty players. He cannot distinguish between the faulty players and the non-faulty players in N.

An equilibrium notion

We now introduce an equilibrium notion for a strategic game in the presence of at most k faulty players. An equilibrium is viewed as a suggested course of action for each player in the game,

that is stable in the following sense. Given an instance (p, k), each non-faulty player has no incentive to deviate from the action suggested to him, regardless of the identity and actions of the faulty players, as long as there are $n - k - 1$ non-faulty players who adhere to the actions suggested to them.

Definition 1. A k-FTNE for the instance (p, k) is a profile of actions $a^* \in A$ having the property that $\forall i \in N$

$$g(a_i^*, a^*_{N\setminus M\cup\{i\}}, a_M) \succsim_i g(a_i, a^*_{N\setminus M\cup\{i\}}, a_M)$$

$\forall a_i \in A_i, \forall a_M \in A_M$ and $\forall M \subseteq N$ such that $|M| \leq k$.

We let $E^k(G, p)$ denote the set of k-fault tolerant Nash equilibria of the game (G, p).

To understand the intuition underlying this equilibrium notion, consider the following scenario. Imagine that the agents about to participate in a mechanism all sit in separate rooms, each in front of his own computer screen. The agents cannot communicate with one another and can only receive messages from the planner. The planner e-mails the agents a common message explaining the rules of the game they are about to play and also suggests the course of action that each agent should take. Suppose that each agent knows that there could be up to k players, who might not have received the message or might not have paid close attention to it or might have not fully understood it (these are the so-called *faulty* players). Each non-faulty player then concludes that up to k players might not follow the planner's advice. However, if the profile of actions that the planner suggested is a k-FTNE, then each non-faulty player will arrive at the following conclusion. As long as all the other non-faulty players (who ever they might be) decide to act according to the planner's advice, I should also follow his advise regardless of who the faulty players are, what they decide to do and how many of them there are.

Implicit in the notion of k-FTNE are the following considerations. Different players may entertain different beliefs regarding the faultiness of others (who the faulty players are, how many of them there are and what actions they choose). Our solution concept requires that a player have no incentive to deviate from his equilibrium strategy for *any* beliefs over the faultiness of players (given that there are at most k faulty players). A less demanding solution concept would imply that the planner knows the beliefs held by each non-faulty player over the faultiness of others.

Potentially any player may be non-faulty. A strategy profile which constitutes a k-FTNE assigns to each player a strategy with the following property. Any player who turns out to be non-faulty has no incentive to deviate from his assigned strategy. That is, our equilibrium notion does not assume that all players are non-faulty. It provides a list of stable strategies for *any* subset of $n - k$ non-faulty players. Thus, if a planner believes that non-faulty players choose their k-FTNE strategies, he is not required to know who the non-faulty players are.

Note that for $k = 0$ we require that the 0-FTNE be a Nash equilibrium (NE), while for $k = n-1$ we require that the $(n-1)$-FTNE be a weakly dominant strategy equilibrium (WDSE). Thus, for $0 < k < n - 1$, the k-FTNE lies between NE and WDSE.

Implementation

Suppose a planner wishes to implement some choice rule f. Let $f(p)$ be interpreted as the set of appropriate outcomes for the profile p. In the standard literature, this planner provides the agents with a game form having the following property. For every possible profile of the agents' preferences, the set of outcomes associated with every equilibrium action profile is identical to the set of outcomes dictated by the choice rule. Assuming that the agents play according to a particular equilibrium notion, the planner wants to be assured that no matter what the current

profile of preferences is and regardless of the specific equilibrium that is selected, the desirable outcome (according to the choice rule) will be realized. In our framework, the planner has the same objective as in the standard framework, only now he is aware that some of the agents might be faulty.

Suppose that some of the players are faulty. Even if all the "good" players satisfy the conditions guaranteeing that they will play their equilibrium strategies, the faulty players will play otherwise. While we have made sure that for every profile of preferences the outcome of every equilibrium action profile is also the one selected by the choice rule, this may not be true for an action profile in which some of the actions are not the equilibrium actions. If the planner believes that non-faulty players will play according to their equilibrium strategies, and he is interested in matching an outcome to a profile of preferences according to some choice rule, the standard notion of implementation is clearly not applicable.

Consider a planner who sets up a second price auction to WDSE-implement the choice rule f that assigns a good to the player with the highest valuation. Since bidding one's true valuation is a weakly dominant strategy,[1] it is also a k-FTNE strategy for any $0 \leq k \leq n-1$. However, even if all the non-faulty players bid their true valuation (their equilibrium strategy), the action of a single faulty player bidding above the true maximal valuation is enough to cause the good not to be sold to the highest valuation player (that is, the good will not be allocated according to the desired choice rule). Note that the notion of implementation in WDSE relies on the assumption that all the players carry out their equilibrium strategies, whereas our notion of implementation requires that we consider the possibility that some of the players might deviate from their equilibrium strategies.

The above example demonstrates that it is not sufficient for a notion of fault tolerant implementation to require only that the solution concept it relies upon be robust to deviations by some players. It must also require that the outcomes of a mechanism be robust to deviations by some players.

Let $a, a' \in A$ be a pair of action profiles. We measure the *difference* between any pair of action profiles, $d(a, a')$, by the number of players who do not choose the same action in both profiles:

$$d(a, a') = |\{i \in N : a_i \neq a'_i\}|.$$

For any profile of actions $a \in A$, we let $B(a, k)$ denote the set of profiles that are different from a by not more than k actions:

$$B(a, k) = \{a' \in A : d(a, a') \leq k\}.$$

We shall refer to $B(a, k)$ as the "k-neighborhood of a".

Definition 2. Let $\langle N, C, P\rangle$ be an environment. The strategic game form G with the outcome function $g : A \to C$ is said to k-FTNE implement the choice rule $f : P \to 2^C$, if $\forall p \in P$, we have

$$g(E^k(G, p)) = f(p) \qquad \text{and} \qquad g(B(a^*, k)) \subseteq f(p)$$

for every $a^* \in E^k(G, p)$.

That is, a mechanism k-FTNE implements a choice rule if it assigns a "desirable" outcome to any action profile in which at least $n - k$ players choose their k-FTNE action. Furthermore,

1. This is true for a setting with asymmetric information, whereas we are dealing with a symmetric information setting. The example is given only to illustrate why the standard definition of full implementation needs to be amended in a setting with faulty players.

any outcome that the choice rule associates with a profile of preferences can be achieved as the result of some k-FTNE action profile.

The multiple equilibrium problem

Our "worst-case-scenario" approach, which requires a mechanism to be robust to any beliefs that players may have about the faulty players, has the cost that fault tolerant mechanisms may have "undesirable" NE, which are not k-FTNE. Thus, if all the players are non-faulty and each plays according to an undesirable NE and each believes that all the other players are non-faulty as well, then no player would have an incentive to deviate; the mechanism would then result in an undesirable outcome. We therefore face the task of explaining the sense in which undesirable NE are unstable.

To appreciate the difficulty of this task consider some undesirable NE. On the one hand, there is *some* belief that player i can have that makes him want to deviate, but on the other hand, there are *other* beliefs that he *could* have that would *not* make him want to deviate. This paper endorses the view that undesirable NE are unstable. A natural criticism of this view is that it considers some beliefs about the faulty players to be more reasonable than others, or, alternatively, this view assumes the planner knows which beliefs a player would hold.

In light of this criticism we offer a possible justification of why NE which are not k-FTNE may not be reasonable. In order to play an equilibrium, which is not k-FTNE, each non-faulty player would need to know the exact predictions the other non-faulty players are making about the faulty players. That is, they must know that those predictions are exactly those that would support the non-k-FTNE. However, since such predictions about other people's beliefs are extremely difficult to make, the players may tend to coordinate instead on k-FTNE, which does not require them to make any predictions (including about other players' predictions).

One possible solution to the multiple equilibrium problem is to consider the following implementation notion: a game form implements a choice rule f if every NE is desirable and every desirable outcome can be obtained as a k-FTNE. Thus, according to this alternative notion there cannot be any undesirable NE; however, there may very well be undesirable k-FTNE. We do not pursue this direction in the current paper. This is a topic we leave for future research.

4. NECESSARY AND SUFFICIENT CONDITIONS

Sufficient conditions

We now present the necessary and sufficient conditions that characterize the family of choice rules that can be k-FTNE implemented in a symmetric information setting, as long as the number of faulty players is not a majority, in other words it is below $\frac{1}{2}n - 1$.

Definition 3. A choice rule $f : P \to 2^C \setminus \{\phi\}$ is k-*monotonic* (k-MON) if whenever $c \in f(p)$, $c \notin f(p')$, then $\exists M \subset N$ and $\exists b \in C$ such that $|M| \geq k + 1$, every $i \in M$ satisfies $c \succsim_i b$, and at least one player $j \in M$ satisfies $b \succ'_j c$.

A choice rule that satisfies k-MON exhibits the following property. If a formerly chosen outcome is excluded from the set assigned to a new profile of preferences, then there are more than k players, each of whom previously considered the chosen outcome to be at least as good as some given outcome, but according to the new profile at least one of these players reverses this relation.

The following example demonstrates that the Core correspondence for a pure exchange economy is at least 1-monotonic.

Example 1. Consider a pure exchange economy in which the set of consequences C is given by $\{x \in \Re_+^{LN} : \sum_i x^i \leq \omega\}$, where ω is an aggregate endowment of L commodities such that each $\omega_i \in \Re_+^L$ and every component of $\sum_i \omega_i$ is positive. Agent i's preferences depend only on his own consumption bundle and are convex, non-decreasing and continuous in this bundle. For a given set of players $S \subseteq N$ let c^S denote an outcome "attainable by S", that is, an outcome in C which satisfies $\sum_{i\in S} x_i = \sum_{i\in S} \omega_i$. The Core is a correspondence that assigns to every profile of preferences p the following set of outcomes:

$$\text{CORE}(p) = \{c \in C : \nexists S \subseteq N \text{ and } \nexists c^S \in C \text{ such that } c^S \succ_i c \text{ for all } i \in S\}.$$

We now show that the Core is at least 1-monotonic. Let $c \in C$ and $p, p' \in P$ such that $c \in \text{CORE}(p)$ but $c \notin \text{CORE}(p')$. Then $\exists S \subseteq N$ and $\exists c^S \in C$ such that $c^S \succ_i' c$ for all $i \in S$. Consider some $j \in S$. By our assumptions on the agents' preferences, there exists $b \in C$ such that $b \sim_j c$ and $b \succ_j' c$. Since $c \in \text{CORE}(p)$, there must be at least one other agent $i \neq j$ who satisfies $c \succsim_i b$. Thus, there is an outcome b and at least two agents who weakly prefer c to b in p, but at least one of them reverses his preferences in p'.

It is instructive to compare k-monotonicity with the notion of monotonicity defined by Maskin (1999).[2] Monotonicity means that if an outcome chosen by a choice rule moves up *everyone's* rankings, then it should continue to be chosen. This property implies that whenever the choice rule excludes a previously chosen outcome, then at least one player has moved this outcome down his ranking. That is, if we denote the formerly chosen outcome by c, then at least one player, who has previously preferred (weakly) c to b now reverses this relation. This does not imply that at least k other players have also preferred (weakly) c to b in the former preference profile. Therefore, as the next example shows, monotonicity does not imply k-monotonicity. However, as shown in Observation 1 below, k-monotonicity does imply monotonicity.

Example 2. $N = \{1, 2, 3\}$, $C = \{a, b, c\}$, $k = 1$, P is the set of all strict preferences over C, and f is a choice rule that chooses an outcome only if it is ranked at the top by some player. f is monotonic, but it is not 1-MON. To see why, assume that f is 1-MON and consider the following pair of preference profiles:

$$\begin{array}{ccc} p_1 & p_2 & p_3 \\ \hline a & b & b \\ b & a & a \\ c & c & c \end{array}, \qquad \begin{array}{ccc} p_1' & p_2' & p_3' \\ \hline b & b & b \\ a & a & a \\ c & c & c \end{array}$$

$f(p) = \{a, b\}$ and $f(p') = \{b\}$. Thus, $a \notin f(p')$. The only difference between p and p' is that player 1 preferred a to b in the former but prefers b to a in the latter. Since f is 1-MON, then it must be that at least one other player ranked a above b in the preference profile p. However, in p both players 2 and 3 prefer b to a, a contradiction. ‖

Observation 1. k-monotonicity implies monotonicity.

Proof. If f is k-monotonic for any $k \geq 1$, then it is also 0-monotonic, and 0-monotonicity is equivalent to monotonicity. ‖

Monotonicity is required for Nash implementation to eliminate unwanted situations in which all the players coordinate on a non-truthful profile of preferences and the resulting outcome

2. A choice rule $f : P \to 2^C \setminus \{\phi\}$ is *monotone* if whenever $c \in f(p)$ but $c \notin f(p')$ there is some player $i \in N$ and some outcome $b \in C$ such that $c \succsim_i b$ but $b \succ_i' c$.

is not the one specified by the choice rule. Suppose that the true profile is p but that all the players coordinate on $p' \neq p$, such that the outcome selected by the mechanism is $c' \notin f(p)$. If f is *monotonic*, then there is always a player who will want the mechanism to take the profile p as input and who will select an outcome which he prefers to c', given this profile.

Monotonicity enables us to construct a mechanism that rules out *Nash* equilibria with non-truthful coordination. However, in some situations, such a mechanism will enable a single player to determine the outcome. This feature might be problematic if we allow some of the players to be faulty. To eliminate non-truthful coordination we still need a preference reversal for at least one player when moving from one profile, p, to another, p', where $f(p) \neq f(p')$. In addition, we need to make sure that if a player challenges the majority view, he will affect the outcome only in those cases in which he is truthful. This can be achieved if we allow only a minority having at least one non-faulty player to affect the outcome.

k-monotonicity, by itself, is not sufficient for k-FTNE implementation. For sufficiency we also require the following property.

Definition 4. A choice rule $f : P \to 2^C \setminus \{\phi\}$ satisfies no-veto-power if $c \in f(p)$ whenever, for at least $n-1$ players, we have $c \succsim c'$ for all $c' \in C$.

Maskin (1999) showed that any choice rule that is monotonic and satisfies no-veto-power is Nash implementable. Similarly, the next result provides sufficient conditions for a choice rule to be k-FTNE implementable.

Proposition 1. *If $n \geq 3$ and $k < \frac{1}{2}n - 1$, then any choice rule that is k-MON and satisfies k-no-veto-power is implementable in k-FTNE.*

Proposition 1 is proved by showing an example of a mechanism, which implements in k-FTNE any k-monotonic choice rule that satisfies no-veto-power. The mechanism we use in our proof is as follows. Each player simultaneously announces a triple (p, c, x), where $p \in P$, $c \in C$ and x is an integer. The outcome function $g : A \to C$ is defined as follows:

Rule 1. If at least $n-k$ players announce (p, c, x) such that $c \in f(p)$, then the outcome is c.

Rule 2. If exactly $n-k-1$ players announce (p, c, x) such that $c \in f(p)$, then the outcome is c, *unless* all of the remaining $k+1$ players agree on $(p', c', \cdot)$ and for everyone of them $c \succsim_i c'$, in which case the outcome is c'.

Rule 3. Otherwise $g((p_i, c_i, x_i)_{i \in N}) = c_j$, where j is such that $x_j \geq x_i$ for all $i \in N$ (in case of a tie the identity of j is immaterial).

The intuition behind the above mechanism is the following. First, if all the non-faulty players coordinate on a "desirable" consequence (that is, a consequence, which the choice rule associates with the true preference profile), then we would like this consequence to be the outcome regardless of how the faulty players behave. Since a group with at least $n-k$ players may consist of only non-faulty players, Rule 1 allows a consensus reached by at least $n-k$ players to determine the outcome.

Rule 2 exploits the k-monotonicity of the choice rule to eliminate "bad" equilibria, in which all players coordinate on a non-truthful announcement. Given a preference profile, k-monotonicity guarantees the existence of a player, who prefers some consequence to one which is not desirable. Thus, if the mechanism allows a majority to determine the outcome, then whenever the majority coordinates on a non-desirable consequence, there exists a player who prefers a different consequence.

There are two obstacles that we need to overcome. First, we need to ensure that when a consensus reached by the majority of the players is being challenged, the players who disagree with the majority include at least one non-faulty player. This is why we require that only a group of $k+1$ players may be able to affect the outcome. Second, whenever the majority agrees on a consequence, the outcome should be determined by the minority if and only if the minority, and not the majority, is being truthful. This requirement is fulfilled by imposing the restriction that all $k+1$ players (the minority) must agree on a consequence, which they prefer less than the consequence proposed by the majority according to the preference profile announced by the majority. Thus, if the majority announces the true preference profile, then no player has an incentive to affect the outcome.

When the majority of the players disagree among themselves, we have no means of verifying who is truthful and who is not. Therefore, we would like to prevent action profiles with disagreement that lead to undesirable outcomes from being equilibria. Whenever the majority of players choose different actions, Rule 3 guarantees that there exists a player with an incentive to deviate.

Taken together, Proposition 1 and Observation 1 imply that any choice rule which is k-FTNE implementable, is also Nash implementable. As k-FTNE reduces to NE when $k = 0$ (there can be no faulty players), Maskin's mechanism for Nash implementation can be obtained by letting $k = 0$ in the above mechanism. The difference between the two mechanisms lies in the fact that when some players may be faulty, we cannot allow the outcome to be determined in equilibrium by a single agent. Since in equilibrium an outcome can only be determined by either Rules 1 or 2, these are the only rules that change when we assume that all players are non-faulty.

We now turn to the formal proof of Proposition 1.

Proof of Proposition 1. If $k = 0$ then k-FTNE implementation reduces to *Nash* implementation. We therefore concentrate on the case where $k > 0$ (since k is an upper bound on the number of faulty players, the actual number of faulty players might still be nil).

We proceed in two steps. The first step establishes that the set of desirable outcomes is contained within the set of k-FTNE outcomes. We show that for any outcome c assigned by the choice rule to a profile of preferences there exists a k-FTNE having the following properties. The equilibrium outcome and also the outcome of any action profile, which is different from the equilibrium by at most k actions, is c.

The second step establishes that for any profile of preferences, every k-FTNE satisfies the following. First, the outcome associated with the equilibrium is desirable (that is, an outcome which the choice rule associates with the profile of preferences). Second, the outcome of any action profile, in which no more than k players deviate from the equilibrium, is also desirable.

Proof of Step 1: Let $c \in f(p)$ for some $p \in P$. Let $a_i = (p, c, 0)$ for each $i \in N$. For every $\hat{a} \in B(a, k)$ we have $g(\hat{a}) = c$. Then (a_i) is a k-FTNE of the game $\langle G, p\rangle$ with the outcome c: a deviation by player j from $\hat{a}$ will cause the outcome to change from c to some c' only if $c \succsim_j c'$.

Proof of Step 2: Let a^* be a k-FTNE of the game $\langle G, p\rangle$ with the outcome c^*. We show that $\forall a \in B(a^*, k)\ g(a) \in f(p)$.

We partition the set equilibria into two separate categories. What distinguishes between the two cases is whether or not the equilibrium satisfies the following property:

($P1$): *Some player can have beliefs about the faulty players that allow him to win an integer game.*

The first category consists of equilibria that do not satisfy this property. These equilibria are characterized by unanimous agreement on a profile of preferences (not necessarily the true one) and a desirable outcome associated with that profile. Unanimity implies that any deviation by k

players or less does not affect the outcome. The k-monotonicity of the choice rule allows us to show that this outcome is also desirable given the true profile of preferences.

The second category consists of all the equilibria which satisfy $P1$. What is special about this class of equilibria is that they all satisfy the following additional property:

($P2$): *If a single player can have beliefs about the faulty players that allow him to win an integer game, then* $n-2$ *other players can also have similar beliefs.*

By using $P2$ we can show that for any deviation of faulty players from the equilibrium, the resulting outcome is the most preferred outcome for at least $n-1$ players. We can then use the k-no-veto-power property of the choice rule to conclude that for any deviation of faulty players, the resulting outcome is desirable.

In order to prove this result we distinguish between two subcases on the basis of whether or not faulty players can trigger the integer game of Rule 3. This distinction is helpful because of the following observation: a belief about the faulty players, which involves an integer game being played, implies that every player can enforce his most preferred outcome by deviating from his equilibrium action. We can then use this observation to prove that every player must satisfy the following property:

($P3$): *Any outcome resulting from an action profile in the k-neighborhood of the equilibrium is at least as good as the player's most preferred outcome.* When faulty players *cannot* trigger an integer game, this property may not hold for one of the players. Still, as long as $n-1$ players satisfy $P3$, then no-veto-power implies that all action profiles in the k-neighborhood of the equilibrium lead to desirable outcomes.

We now turn to the formal proof of the cases discussed above.

Case 1. Suppose there exists no $a \in B(a^*, k+1)$ such that $g(a)$ is determined by Rule 3. This means that the equilibrium a^* is such that for all $i \in N$, $a_i^* = (p^*, c^*, \cdot)$ with $c^* \in f(p^*)$. Assume that $c^* \notin f(p)$. Since f is k-MON, there is a subset M of at least $k+1$ players and a consequence $b \in C$ that satisfy $c^* \succsim_i' b$ for all $i \in M$ and $\exists j \in M$ such that $b \succ_j c^*$. It follows that for a profile $a' \in A$ in which k players in the subset $M\backslash\{j\}$ play $(p, b, \cdot)$ and the rest of the players play a^*, player j's best response is $(p, b, \cdot)$ and not a_j^*. Since this contradicts our initial assumption that $a^* \in E^k(G, p)$, we conclude that $c^* \in f(p)$.

Suppose $g(a) = c^*$ for every $a \in B(a^*, k)$. This means that the equilibrium outcome c^* cannot be affected by *any* deviating minority. Thus, there exists an agent j such that for all $i \in N\backslash\{j\}$ we have $a_i^* = (p^*, c^*, \cdot)$ with $c^* \in f(p^*)$, and agent j either announces the outcome c^* or an outcome c that satisfies $c \succ_j^* c^*$.

Case 2. Suppose Rule 3 applies to some $a \in B(a^*, k+1)$. We show that in this case, (1) at least $n-1$ players weakly prefer c^* to any other outcome, and (2) every player is indifferent between c^* and any $g(a)$ with $a \in B(a^*, k)$. This implies that for any $g(a)$ with $a \in B(a^*, k)$, at least $n-1$ players weakly prefer $g(a)$ to any other outcome. Thus, by no-veto-power, $g(a) \in f(p)$ for any $a \in B(a^*, k)$.

We proceed by considering each of the following cases separately.

Case A. There exists some $a' \in B(a^*, k)$ such that $g(a')$ is determined by Rule 3, and $g(a') \neq c^*$. Let $M \subset N$ be the subset of players who deviate from a^* in the action profile a'; that is, $M = \{i \in N : a_i' \neq a_i^*\}$. Let x be the largest integer announced in the profile a'. Let $\tilde{a}$ be an action profile in which all the players in M coordinate on some $c \in C$ and on an integer $y > x$ while $\tilde{a}_i = a_i^*$ for all $i \in N\backslash M$. Then $\tilde{a} \in B(a^*, k)$ and $g(\tilde{a}) = c$ by Rule 3. Let $\hat{a}$ be an action profile in which some $j \in N\backslash M$ announces $\hat{a}_j = (\cdot, b, z)$ where $z > y$ and $\hat{a}_{-j} = \tilde{a}_{-j}$. By Rule 3, $g(\hat{a}) = b$. Thus, for any $i \in N$, there is some belief about the faulty players that would make him want to deviate: If i expects $\tilde{a}_{-i}$ to be played, then he can profit by announcing his most favorite outcome and an integer $y > z$. Since by assumption a^*

is a k-FTNE, no player can have an incentive to deviate. It follows that every player must be indifferent among all outcomes.

Case B. There exists no $a' \in B(a^*, k)$ such that $g(a')$ is determined by Rule 3, and $g(a') \neq c^*$.

Claim 1. *For all $a \in B(a^*, k)$, $c^* \backsim_i g(a)$ for all players.*

Proof of Claim 1. Note that Case B can only occur if the equilibrium a^* satisfies the following: $\exists j \in N$ such that $\forall i \in N\backslash\{j\}$ we have $a_i^* = (p^*, c^*, \cdot)$ with $c^* \in f(p^*)$ while $a_j^* = (\cdot, c^j, \cdot)$ with $c^j \neq c^*$. It follows that any $a \in B(a^*, k)$ with $g(a) \neq c^*$ must satisfy $g(a) = c^j$.

Consider some $a \in B(a^*, k)$ with $g(a) \neq c^*$. Let S be the subset of k players other than j who choose the consequence c^j in a. Consider the $n - k - 1$ players belonging to $N\backslash(S \cup \{j\})$. Each of these players must be indifferent between c^* and c^j. To see why, consider some player $i \in N\backslash(S \cup \{j\})$. Let x be the largest integer announced in a^*. Suppose player i believes that j is non-faulty, but that the members of S are faulty. Suppose further that i believes that these k faulty players deviate from their equilibrium action to a_j. Given this belief about the faulty players, player i can either enforce the outcome c^* by deviating from a_i^* to (p^*, c^*, y) where $y > x$, or he can decide not to deviate and have c^j as the outcome. Since a^* is a k-FTNE, player i must be indifferent between c^* and c^j. It follows that for all $i \in N\backslash(S \cup \{j\})$ and for all $a \in B(a^*, k)$, $c^* \backsim_i g(a)$. To see that each member in S is indifferent consider a deviation by all other members in S to a_j.

We now turn to consider player j. If $c^j \succ_j^* c^*$, then $g(a) = c^*$ for all $a \in B(a^*, k)$. Assume $c^* \succsim_j^* c^j$. Then there is a belief that player j can hold about the faulty players that would make him want to deviate: if k players other than him coordinate on a_j^*, then player j would want to choose the outcome c^*. ‖

Claim 2. *At least $n - 1$ players weakly prefer c^* to any other outcome.*

Proof of Claim 2. Consider some $i \in N\backslash\{j\}$. By our assumption that $n \geq 3$ and that $k < \frac{1}{2}n - 1$, there exists an action profile $\tilde{a} \in B(a^*, k)$ in which k players in $N\backslash(\{j\} \cup \{i\})$ announce the triplet $(p', c', 0)$ where $p' \neq p^*$ and $c' \neq c_j$. Suppose i's belief about the faulty players is such that he expects $\tilde{a}_{-i}$ to be played· Given this belief about the faulty players, player i can enforce his most favorite outcome c^i by deviating from a_i^* to $a_i^{**} = (p^i, c^i, y)$ where $p^i \neq p^*$ and $y > x$. Since a^* is a k-FTNE, player i has no incentive to deviate from a^*. Thus, for all $i \in N\backslash\{j\}$ and $c \in C$ we have $c^* \succsim_i c$.

Note that player j, unlike the other players, may not be able to achieve his favorite outcome by means of this deviation since all the other players might be coordinating on the triple $(p, c^*, 0)$, where p is the true profile of preferences. ‖

By Claim 2 and the no-veto-power of f, $c^* \in f(p)$. By Claim 1, there are at least $n - 1$ players who satisfy $g(a) \succsim_i c$ for all $c \in C$ and $a \in B(a^*, k)$. By the no-veto-power of f, $g(a) \in f(p)$ for all $a \in B(a^*, k)$. ‖

A necessary condition

If a social choice *function* satisfies no-veto-power, then k-MON is both necessary and sufficient for k-FTNE implementation. However, a social choice *correspondence* which is implementable in k-FTNE satisfies a weaker notion of k-monotonicity.

Definition 5. A choice rule $f : P \to 2^C \backslash \{\phi\}$ is *weakly k-monotonic* if whenever $f(p) \nsubseteq f(p')$, then $\exists M \subseteq N$ having at least $k+1$ players and $\exists b \in C$ such that for every player $i \in M$, there is an outcome $c^i \in f(p)$ satisfying $c^i \succsim_i b$, and for at least one of these players, say j, $b \succ'_j c^j$.

A choice rule that satisfies weak k-MON exhibits the following property. Whenever a formerly chosen outcome is excluded from the set associated with a new profile of preferences, then there are least $k+1$ players, each of whom previously considered one of the chosen outcomes to be at least as good as some given outcome, but according to the new profile at least one of these players reversed this relation.

The main difference between the two notions of k-monotonicity is the following. A choice rule which satisfies k-MON excludes a formerly chosen outcome from a newly chosen set, only if the ranking of *this* outcome has changed. On the other hand, a weakly k-MON choice rule may exclude a formerly chosen outcome from a newly chosen set, even if only the rankings of *other* chosen outcomes have changed. Thus, if a choice rule is k-MON, then it is also weakly k-MON. In addition, for choice *functions* the two notions coincide.

The next proposition shows that a social choice *correspondence* which is implementable in k-FTNE must satisfy weak k-monotonicity.

Proposition 2. *If a choice rule is k-FTNE implementable, then it is weakly k-MON.*

Before we proceed with the proof of Proposition 2 we provide an intuitive explanation for why an implementable choice correspondence need not be k-MON. Consider an outcome c^* which is chosen for the profile p but not for p'. This implies that c^* is an equilibrium outcome for p but not for p'. Thus, the action profile that results in c^* (say a^*) is stable for p but not for p'. It follows that some player i has an incentive to deviate from a_i^*. However, this incentive to deviate is not necessarily a result of a preference reversal between c^* and some other outcome b (such that $g(a^*) = c^*$ and $g(a_i, a^*_{-i}) = b$). There might be some deviation of up to k players that makes a deviation from a_i^* profitable when the profile of preferences is p'.

Let a'_S denote a deviation of a subset of no more than k players. The change from p to p' has led player i to reverse his preferences between $g(a_i^*, a'_S, a^*_{N\backslash(\{i\}\cup S)})$ and $g(a'_i, a'_S, a^*_{N\backslash(\{i\}\cup S)})$. Let $c^i \equiv g(a_i^*, a'_S, a^*_{N\backslash(\{i\}\cup S)})$. If the choice rule is implementable, then c^i is a member of the chosen set for p, but it may be different from c^*. Thus, c^* may have been excluded from the set of outcomes chosen for p' because of a preference reversal between outcomes other than c^*.

Proof of Proposition 2. Let $f : P \to 2^C \backslash \{\phi\}$ be a choice rule that is k-FTNE implementable by a strategic game form $G = \langle N, (A_i), g\rangle$ and let $p, p' \in P$ such that $f(p) \nsubseteq f(p')$. It follows that $\exists c^* \in C$ which is an element in $f(p)$ but not an element in $f(p')$. Since G implements f in k-FTNE there must be an action profile $a^* \in A$ satisfying $g(a^*) = c^*$, $a^* \in E^k(G, p)$ and $a^* \notin E^k(G, p')$.

If $a^* \notin E^k(G, p')$ then there is a subset $S \subseteq N$ with at most k players, a player $j \in N\backslash S$ and an action profile $(\hat{a}_j, \hat{a}_S) \in A_j \times A_S$ such that

$$g(\hat{a}_j, \hat{a}_S, a^*_{N\backslash(S\cup\{j\})}) \succ'_j g(a_j^*, \hat{a}_S, a^*_{N\backslash(S\cup\{j\})}). \tag{1}$$

Since it might be the case that $|S| < k$ we let $S^k \subseteq N\backslash\{j\}$ such that $S \subseteq S^k$, $|S^k| = k$, and $\hat{a}_{S^k} = ((\hat{a}_i)_{i\in S}, (a_i^*)_{i\in S^k\backslash S})$. In order to conform to the notations used in the definition of weak k-MON we let

$$M \equiv S^k \cup \{j\}$$

$$b \equiv g(\hat{a}_j, \hat{a}_{M\setminus\{j\}}, a^*_{N\setminus M}) \qquad \text{and}$$
$$c^i \equiv g(a^*_i, \hat{a}_{M\setminus\{i\}}, a^*_{N\setminus M}) \qquad \text{for any } i \in N.$$

If equation (1) holds, then $b \succ'_j c^j$. Since $a^* \in E^k(G, p)$, then every $i \in M$ satisfies $c^i \succsim_i b$.

From the definition of $B(a^*, k)$ it follows that

$$c^i \in B(a^*, k) \qquad \text{for all } i \in M.$$

Since G implements f in k-FTNE, we have

$$c^i \in f(p) \qquad \text{for all } i \in M.$$

It follows that f is weakly k-MON. ‖

Maskin (1999) showed that any Nash-implementable choice rule must be monotonic. The set of monotonic choice rules and the set of weakly k-MON choice rules are two distinct sets. As the next three examples demonstrate, the two notions of monotonicity have a non-empty intersection which includes the weakly Pareto efficient choice rule (Example 3). However, there are weakly k-MON which are not monotonic (Example 4) and there are monotonic choice rules which are not weakly k-MON (Example 5).

Example 3. A choice rule f associates the set of weakly Pareto efficient outcomes in C with a preference profile p if it satisfies

$$f(p) = \{c \in C : \nexists b \in C \text{ such that } b \succ_i c \ \ \forall i \in N\}.$$

We now proceed to show that for $k < n$ any Pareto efficient choice rule f satisfies weak k-MON. Let $p, p' \in P$ and let $c^* \in C$, such that $c^* \in f(p)$ but $c^* \notin f(p')$. Thus, $\exists b \in C$ such that $\forall i \in N$, $b \succ'_i c^*$. However, since $c^* \in f(p)$, $\exists j \in N$ who satisfies $c^* \succsim_j b$. Let $\overline{c}_i(p)$ denote the most weakly preferred outcome for agent i when the profile of preferences is p (that is, $\overline{c}_i(p) \succsim_i c \ \forall c \in C$). Thus, $\forall i \in N \ \overline{c}_i(p) \succsim_i b$. From the fact that f is a Pareto efficient choice rule it follows that $\{\overline{c}_i(p)\}_{i\in N} \subseteq f(p)$.

Therefore, there is a set of $n \geq k+1$ players (the set N), each of whom has an element in $f(p)$ (c^* for player j and $\overline{c}_i(p)$ for every $i \neq j$) that he weakly prefers to the outcome b when the preference profile is p. However, when the preference profile is p', at least one of these players (player j) reverses his preference.

Example 4. $N = \{1, 2, 3\}$, $C = \{a, b, c, d\}$, $k = 1$ and P is the set of all strict preferences over C. The choice rule f chooses d only if it is ranked at the top by at least two players. Otherwise, the chosen set includes any outcome which some player prefers to d. We now check that f is weakly k-MON. Suppose $f(p') = \{d\}$ but $d \notin f(p)$. It follows that by p' at least two players preferred d to any other outcome, but by p at least one of them now ranks a different outcome at the top. It follows that f is weakly k-MON. However, f is not monotonic. To see why, consider the following pair of preference profiles:

p_1	p_2	p_3		p'_1	p'_2	p'_3
b	b	b		d	d	a
d	d	c	,	a	a	b
c	c	a		b	b	c
a	a	d		c	c	d

$f(p) = \{a, b, c\}$ and $f(p') = \{d\}$. Although the outcome a has moved up everyone's rankings, it is excluded from $f(p')$. It follows that f is not monotonic.

Example 5. Consider a finite set of consequences C. Assume each player has a strict preference relation over the elements in this set. Let P consist of all the possible strict preference profiles. A choice rule f is dictatorial if it satisfies $f(p) = \{c \in C : c \succsim_j b \ \forall b \in C\}$, where j is some player (the dictator) in N. While a dictatorial choice rule is Nash-implementable, it is easy to see that it violates weak k-MON for any positive k.

With regards to the last example, note that a choice rule, which is implementable in Nash but not in k-FTNE, must satisfy the following: there is at least one preference profile in which the outcome c is chosen even though $n-1$ players agree that a different outcome dominates c. It follows that the necessary condition for k-FTNE implementation is not violated by a monotonic choice rule which does not select an outcome, if $n-1$ players agree that a different outcome is better.

5. EXAMPLES OF MECHANISMS

The constrained Walrasian function

The canonical mechanism we used in the proof of the necessary and sufficient conditions was abstract and relied on the usage of integer games. For more concrete settings, such as exchange economies, we will use a more specific game form that resembles some kind of trading mechanism. In this section we present a simple mechanism that k-FTNE implements the constrained Walrasian function.

Let $E = \langle N, \omega, p\rangle$ be a pure exchange economy with free disposal, where p is an element in the set P of all continuous and strictly convex preference relations (such that for every economy there is a unique constrained Walrasian allocation) and $\omega_i > 0 \ \forall i \in N$. An allocation $x^* \in \Re_+^{Ln}$ is *constrained Walrasian* if it satisfies that $\exists \lambda^* \in \Re_+^L$ with $\lambda^* \neq 0$ (the price vector) such that $\forall i \in N$, $\lambda^* x_i^* = \lambda^* \omega_i$ and $x_i^* \succsim_i x_i$ for all $x_i \leq \sum_{i\in N}(\omega_i)$ such that $\lambda^* x_i \leq \lambda^* \omega_i$. The pair (λ^*, x^*) will be called the *competitive equilibrium* of the exchange economy. The set of constrained Walrasian allocations of an exchange economy E will be denoted by $CW(E)$; the set of all its competitive equilibria will be denoted by $CE(E)$.

Consider the following game form G. Each player $i \in N$ simultaneously announces a pair $a^i = (\lambda^i, x^i)$, where $\lambda^i \in \Re_+^L$ (the vector of prices) and $x^i \in X(E)$ (an allocation of goods).[3] The outcome function $g : \Re_+^L \times \Re_+^{Ln} \to \Re_+^{Ln}$ is defined as follows:

Rule (G1): If at least $n-k$ traders agree on (λ^*, x^*) and $\lambda^* x_i^* \leq \lambda^* \omega_i \quad \forall i \in N$, then x^* is implemented.

Rule (G2): If exactly $n-k-1$ traders agree on (λ^*, x^*) and $\lambda^* x_i^* \leq \lambda^* \omega_i \quad \forall i \in N$, then x^* is still implemented *unless* all the remaining $k+1$ traders announce a pair (λ^*, y) satisfying the following two properties, in which case y is implemented:

- $\lambda^* y_i \leq \lambda^* \omega_i \quad \forall i \in N$.
- For all of these $k+1$ traders $y_i \neq x_i^*$.

Rule (G3): In all other cases $g((\lambda^i, x^i)^{i\in N}) = \hat{x}$, where $\hat{x}_i = \hat{x}_i^M$ if there is a *unique maximal* subset $M \subseteq N$ of at least $k+1$ traders containing i, such that all the members of M agree on the allocation $\hat{x}^M$ that satisfies $\sum_{j\in M}(\hat{x}_j^M) = \sum_{j\in M}(\omega_j)$ and $\hat{x}_j^M \neq \omega_j \quad \forall j \in M$. Otherwise, $\hat{x}_i = \omega_i$. In other words, trade takes place only within the largest mutually disjoint sets of agents who agree to exchange their initial endowments.

3. Note the difference between a superscript and a subscript. For example, x_j^i denotes j's bundle of goods in the allocation proposed by player i.

Proposition 3. *The game form G implements the constrained Walrasian function of the economy E in k-FTNE as long as* $0 < k < \frac{1}{2}n - 1$.

The proposed mechanism has the same general structure as the canonical mechanism of the proposition. It is less abstract since it exploits the special features of the particular environment in which it is set. This is best demonstrated in the last two rules of the mechanism.

Rule G2 has two objectives. First, it relies on the k-monotonicity of the Walrasian function to eliminate "bad" equilibria, in which all players coordinate on a non-truthful announcement. Second, it guarantees that whenever the majority agrees on a feasible allocation, the outcome should be determined by the minority if and only if the minority, and not the majority, is being truthful. This requirement is fulfilled by imposing the restriction that all $k + 1$ players (the minority) agree on an allocation and a price vector which satisfy two conditions; the prices must be identical to the ones announced by the majority; the allocation is different from the one announced by the majority in the bundles of goods assigned to each member of the minority. Since for each profile of preferences there is a unique Walrasian allocation; and since this allocation is optimal for each player given a vector of prices; no player would have an incentive to affect the outcome, if that outcome assigns a Walrasian allocation to the economy.

Rule G3 takes on the role of the integer game used in the proof of Proposition 2. It treats action profiles in which the majority of players disagree and the resulting outcome is undesirable. To prevent such action profiles from being equilibria we need to be able to construct a deviation by k players which would motivate some other player to deviate as well. Using Rule G3 we can find a group of k players that satisfy the following. Either a player, who is not trading in the current action profile, would want to imitate the action of these k players, (if, for example, they deviate from their current action profile and offer him a sufficient amount of goods for free); or a player, who is currently trading, would want to change his action, so that he could still be a member of a set of players that are allowed to trade among themselves (if, for example, a deviation of k players changes the set of players who agree to trade among themselves).

Proof of Proposition 3. See Appendix. ‖

Allocation of an indivisible good

Consider a planner who has to decide whether to allocate an indivisible good to one agent in a group of n or whether to keep the good in his possession. Let I be the set of $n + 1$ possible allocations of the good. For each agent i we denote by v_i the value that he associates with owning the good. Let V be the set of all possible profiles of valuations such that $V \subseteq \Re_+^n$. The profile of valuations is known to the agents but not to the planner.

The n agents have a preference relation defined over $I \times V$. Let $i^*(v)$, $\forall v \in V$, denote the agent whose valuation is the highest. Let (i, v) denote the outcome in which the profile of valuations is $v \in V$ and the good is allocated to agent i. In the allocation $(\emptyset, v)$ the planner keeps the good. $\forall v \in V$ each agent i satisfies $(i, v) \succsim_i (i^*(v), v)$ with strict preference in case $i \neq i^*(v)$. In addition $(\emptyset, v) \succ_i (j, v)$, where $j \in N \backslash (\{i\} \cup \{i^*(v)\})$.

The preferences of the agents can be described as follows. Each agent would like to own the good. However, if an agent does not receive the good, he would like the good to be given to the agent who values it the most. Each agent prefers that the good will not be allocated at all to the allocation in which it falls to the hands of an undeserving agent (any agent besides himself who is not the highest valuation agent). These preferences capture a situation in which the agents care about who owns the good. An example where such preferences may prevail is the allocation of a prize (say, an apartment): each of the participants would like to receive the prize, but if he does

not win, then he would prefer the prize to be given to a needy homeless person (rather than to his next door neighbor).

The planner wishes to k-FTNE implement the following choice rule: $f(v) = (i^*(v), v)$ $\forall v \in V$.[4]

Proposition 4. *If $0 < k < \frac{1}{2}n - 1$, then the following mechanism G implements f in k-FTNE. Each agent i announces a pair $a_i = (j_i, S_i)$, such that $j_i \in N$ and $S_i \subseteq N$.*

Rule 1: *If at least $n - k$ agents agree on agent j, the object is given to him.*

Rule 2: *If exactly $n - k - 1$ agents agree on agent j, then the object still goes to him, unless the remaining subset M of $k + 1$ agents all agree on h such that $h \notin M$, in which case h receives the good.*

Rule 3: *Otherwise, the object remains with the planner unless there is a unique maximal $S \subseteq N$ satisfying:*

1. $|S| \geq k + 1$.
2. *All agents in S agree that agent j should get the good and $j \in S$.*
3. $a_j = (j, S)$.

The innovation of the above mechanism lies in its third rule, which applies to action profiles in which there are at least three different views as to who should receive the good. Given such an action profile, Rule 3 allows us to construct a deviation of k players such that some player would have an incentive to deviate.

Suppose the action profile satisfies that the planner keeps the good for himself. If a group of k players deviate from their current action profile by announcing that the good should be given to player j, who is not a member of this group; player j would want to deviate from his chosen action by declaring that he should receive the good and by naming the k deviators (recall that Rule 3 requires the recipient of the good to name himself and the set of players who chose him). Suppose, on the other hand, that the action profile satisfies that a player i receives the good. Player i would have an incentive to deviate and change the set of players in his announcement, if a player, who has not chosen i as the rightful owner of the good in the current action profile, now decides to do so.

The proof of Proposition 4 is omitted as it follows the same line of argument as the proof of Proposition 3 (a detailed proof may be found in Eliaz, 1999).

6. CONCLUDING REMARKS

In this section we suggest and discuss some interpretations and extensions of our framework. First, we discuss how our framework needs to be amended such that our results will also hold when we interpret faulty players as players whose preferences are unknown. Next, we present an interpretation of our model, in which faulty players are viewed as players who need to learn how to choose optimal actions in a game. An interpretation of our non-Bayesian approach is also discussed; followed by a suggestion as to how our framework could be revised to allow for a Bayesian analysis of fault tolerance. We close with a summary of our main results.

Faulty players as players with unknown preferences

Throughout our analysis we have interpreted the unpredictable behavior of faulty players as the result of mistakes. An alternative interpretation would be that faulty players are agents whose

4. A related problem is the design of auction with externalities, which is discussed in Jehiel *et al.* (1996, 1999).

preferences are unknown. However, we still want to retain the assumption that the exact number and identity of the faulty players are not known. In order for this assumption to be consistent with our alternative interpretation we can assume that all the non-faulty players share the same preferences. That is, if i and j are non-faulty players, then $\forall p \in P$ we have $p_i = p_j = p^*$. Furthermore, player i knows that any player who is non-faulty also has that preference relation p^*, but i cannot identify such a player. In this setup, a player whose actions are inconsistent with p^* (and is thus faulty) might simply have different preferences that are not known. The planner knows that all the non-faulty players have the same preference relation, but he knows neither this relation nor the identity of the non-faulty players (that is, he only knows the structure of the profile of preferences).

In this setup the terms "faulty" and "non-faulty" might not seem appropriate. It is more natural to consider a population in which a minority of "outsiders" or "foreigners" is assimilated among the "natives" or "locals". The planner wishes to associate a set of outcomes with the preference relation of the native population. Although the domain of the choice rule is the set of preference profiles of the natives exclusively, the planner cannot prevent the foreigners from participating and affecting any mechanism that he constructs.

All of our previous results continue to hold in this special set up. Here, instead of announcing an entire profile of preferences, a player would be asked to report a single preference relation. Thus, if in equilibrium, each native truthfully reveals his own preference relation, then regardless of the actions of the foreigners (as long as their number is below half the size of the total population minus one), the desired outcome will be implemented.

Faulty players as players who learn how to play

An underlying assumption in the standard theory of implementation is that the players' behavior is consistent with some game theoretic solution concept. That is, the players execute the equilibrium strategies. Suppose a mechanism is set up and repeatedly played for an indefinite period of time. Suppose further that the players are myopic in the sense that they care only about the outcomes in each constituent game. However, not all the players instantaneously arrive at the equilibrium of the constituent game. Some need to participate in the mechanism for several periods in order to gain experience and eventually learn how to play the game. In such a setting, even if a mechanism fully implements a choice rule in the standard sense, the players may not play the equilibrium strategies right from the start, but only after a significant amount of time.

In an environment where some players learn how to play, the planner needs to worry about the path of play. If we conceive of a faulty player as a player who learns, we can reinterpret our result in a way that fits this dynamic setting. If the number of learners is strictly less than $\frac{1}{2}n - 1$, the planner can guarantee in each and every period of any learning process the full implementation of any k-MON choice rule that satisfies no-veto-power.[5]

5. Recently, several papers attempt to answer the question whether agents who learn how to play a mechanism will actually arrive at the desired equilibrium. Cabrales (1998) studies the equilibrating process of several implementation mechanisms using naive adaptive dynamics and shows that the dynamics converge and are stable for the canonical mechanism of implementation in *Nash* equilibrium. However, for more refined equilibrium concepts he shows that the dynamics converge but are not stable.

Cabrales and Ponti (1998) study the convergence and stability properties of a mechanism that implements most social choice functions using the solution concept of iterated elimination of weakly dominated strategies. They show that if the players learn through monotonic dynamics, their strategies may converge to *Nash* equilibria whose outcomes are different from those desired by the choice rule.

Hon-snir *et al*. (1998) analyze a repeated first-price auction and show that if every player is using either a belief-based learning scheme with bounded recall or a generalized fictitious play learning scheme, then the players' bids will converge to the equilibrium bids of a one-shot auction in which the types are commonly known.

In these papers, however, the convergence may take a considerable amount of time.

A Bayesian approach

Although we follow a non-Bayesian approach one may criticize our assumption that an upper bound k is common knowledge among the planner and the non-faulty agents. This assumption stems from our view of a planner as a professional who is elected/hired by the agents to construct a mechanism that can implement a given choice rule (much like a computer engineer who is hired to construct a computer network). In our framework, when the planner finishes his work he can approach the agents and credibly announces that his mechanism works only if less than $\frac{1}{2}n - 1$ of them make mistakes. If the agents can accept this "fault tolerance", then they should adopt his mechanism.

One consequence of our reliance on common knowledge of an upper bound k is that our model is sensitive to the value of k. A mechanism, which implements a choice rule in k-FTNE may not implement the same choice rule in $(k-1)$-FTNE. Thus, a nice topic for future research would be to find a mechanism that works for k as well as for any $k' < k$.

It would be interesting to investigate whether our results can also be obtained in a somewhat Bayesian framework. The model we have in mind is one where there is a small probability ε that any player will turn out to be faulty. However, it is still not known what a faulty player might do (since it seems difficult to come up with a reasonable support for the set of possible behaviors of a faulty player). The equilibrium and implementation notions then need to be appropriately amended.

Summary

Many institutions and mechanisms rely on intricate sets of rules that are not always fully understood by the individuals who must follow them. Consequently, some of the individuals might fail to choose the optimal course of action. We can therefore ask whether it is possible to design a mechanism that achieves their objective even if some of the participants make mistakes and act in an unpredictable manner.

In this paper we have tried to provide a partial answer to this question. We first developed a theoretical framework for analyzing fault tolerant implementation. This required us to define a new equilibrium notion, k-FTNE, and a new notion of implementability. Using this framework we have identified the necessary and sufficient conditions for implementation in k-FTNE. Since our proof relies on the construction of an abstract mechanism involving integer games, we presented two examples of simple mechanisms. The first example described a trading mechanism, in which participants choose prices and allocations, that k-FTNE implements the constrained $Walrasian$ function. In the second example we present a simple procedure for efficiently allocating an indivisible good.

We have only taken a very small step towards the incorporation of fault tolerance into the theory of implementation. The next step could be the study of fault tolerant implementation via extensive game forms. Extensive form games introduce an additional difficulty: unlike strategic form games if players act in a sequential order, they can update their beliefs regarding the number and identity of the faulty players in the population (for example, if a non-faulty player acts following a history which is not on the equilibrium path). In addition, we have only considered the case of symmetric information among the non-faulty players. A natural extension would be to analyze implementation in *k-fault tolerant Bayesian Nash equilibrium* for settings with asymmetric information.

APPENDIX

Proof of Proposition 3. The method of proof is the same as in Proposition 1. By arguments similar to those used in the proof of Proposition 1 it can be shown that if $x = CW(E)$, $\exists a^* \in E^k(G, p)$ such that $\forall a \in B(a^*, k)$, we have

$g(a) = x$. To complete the proof we need to show that if $a^* \in E^k(G, p)$, then $\forall a \in B(a^*, k)$, $g(a) = CW(E)$. As in the proof of Proposition 1, we consider three different cases.

The first case covers equilibria with unanimous agreement on a price vector and a feasible allocation. It can be proven by using essentially the same arguments of the corresponding case in the proof of Proposition 1.

In the second case, exactly $n-1$ players announce the same pair (λ^*, x^*), which satisfies $\lambda^* x_i^* \leq \lambda^* \omega_i \ \forall i \in N$. We denote the subset of $n-1$ coordinating players by M (such that $\forall i \in M \ \ a_i^* = (\lambda^*, x^*)$) and the n-th player by j (such that $a_j^* = (\lambda', x')$). If $a^* \in E^k(G, p)$, then $x^* = CW(E)$. Otherwise, $(\lambda^*, x^*) \notin CE(E)$. By free disposal, there exists a player $h \in N$ and an allocation y such that $y \succ_h x^*$, $\lambda y_i \leq \lambda \omega_i \ \ \forall i \in N$ and $y_i \neq x_i^* \ \ \forall i \in N$. It then follows that $\exists a \in B(a^*, k)$ such that given a_{-h}, player h would want to deviate from a_h^*, a contradiction (take a in which (λ^*, y) is announced by either player j and $k-1$ players in $M \backslash \{h\}$ if $h \in M$ or by k players in M if $h = j$).

It remains to be shown that every equilibrium a^* covered by Case 2 satisfies that $\forall a \in B(a^*, k)$, $g(a) = x^*$. We show that if $a^* \in E^k(G, p)$ then it must be that the price λ^* announced by the $n-1$ players in M differs from the price λ' chosen by j. Suppose not, so $a_j^* = (\lambda^*, x')$. Suppose $\exists a' \in B(a^*, k)$ such that $g(a') \neq x^*$. Rule G2 of the mechanism implies the following: $a_j^* = (\lambda^*, x')$ where $x'_j \neq x_j^*$, $\lambda^* x'_i \leq \lambda^* \omega_i \ \forall i \in N$, and a' satisfies that k players in M coordinate on (λ^*, x'). From the continuity and strict convexity of p and from our conclusion that $x^* = CW(E)$, it follows that $x^* \succ_j x'$. Thus, given a'_{-j}, player j would prefer to deviate from a_j^* to (λ^*, x^*), a contradiction.

We show that a profile, not covered by the previous two cases (either a unanimous agreement on a price-allocation pair such that at least one player cannot afford his bundle, or at least three players disagree), cannot be a k-FTNE of the game. Suppose that $\exists a^* \in E^k(G, p)$ which is not covered by cases 1 or 2. The outcomes of the action profiles in which k players deviate from a^* are determined according to rule G3 of the mechanism. We construct an action profile $a \in B(a^*, k)$ having the following properties: a set M of k players coordinate on an allocation x in which some player $j \notin M$ receives a bundle x_j that is different from the bundle that he announced in a_j^* and which he strictly prefers to ω_j (for example, by adding goods free of charge to his initial bundle). In addition, x is different from any allocation announced by players outside of M and satisfies $\sum_{i \in M \cup \{j\}}(x_i) = \sum_{i \in M \cup \{j\}}(\omega_i)$. The set M and player j are chosen such that in the outcome $g(a)$, player j does not trade.[6] However, given a_{-j}, if j deviates from a_j^* to the announcement of the players in M, he obtains the bundle x_j. By construction, this deviation is profitable, a contradiction. ‖

Acknowledgements. I would like to thank my adviser, Ariel Rubinstein, for his help, encouragement and support throughout this project. I thank the editor of this journal and two thoughtful referees for helping me substantially improve this paper. I would also like to thank Markus Brunnermeier, Tilman Börgers,Yoram Hamo, Leonardo Felli, Jacob Glazer, Phillipe Jehiel, Matthew Jackson, Michael Ornstein, Michele Piccione, Ilya Segal and Ran Spiegler for their many helpful comments. I thank seminar participants at the University of Bristol, Cambridge, University of Michigan, NYU, LSE, Princeton, UCL, Summer in Tel-Aviv and Games 2000 for their valuable feedback. I appreciate the hospitality extended to me by STICERD at the London School of Economics where part of this research was undertaken. This research was partially supported by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities. Financial Support from the European Commission's Marie Currie Fellowship (TMR grant) is gratefully acknowledged.

REFERENCES

BARTHOLDI, J. J. III, TOVEY, C. A. and TRICK, M. A. (1989), "The Computational Difficulty of Manipulating an Election", *Social Choice and Welfare*, **6**, 227–241.

CABRALES, A. (1999), "Adaptive Dynamics and the Implementation Problem with Complete Information", *Journal of Economic Theory*, **86**, 159–184.

CABRALES, A. and PONTI, G. (2000), "Implementation, Elimination of Weakly Dominated Strategies and Evolutionary Dynamics", *Review of Economic Dynamics*, **3**, 147–282.

ELIAZ, K. (1999), "Fault Tolerant Implementation" (Working Paper 21-99, Foerder Institute for Economic Research, Tel-Aviv University).

HON-SNIR, S., MONDERER, D. and SELA, A. (1998), "A Learning Approach to Auctions", *Journal of Economic Theory*, **82**, 65–88.

HURWICZ, L. (1986), "On the Implementation of Social Choice Rules in Irrational Societies", in W. P. Heller, R. M. Ross and D. A. Starret (eds.) *Social Choice and Public Decision Making: Essays in Honor of Kenneth J. Arrow* (Cambridge: Cambridge University Press).

JEHIEL, P., MOLDOVANU, B. and STACCHETTI, E. (1996), "How (Not) to Sell Nuclear Weapons", *American Economic Review*, **86**, 814–829.

JEHIEL, P., MOLDOVANU, B. and STACCHETTI, E. (1999), "Multidimensional Mechanism Design for Auctions with Externalities", *Journal of Economic Theory*, **85**, 258–293.

6. If in a^* there is no trade then any set of $k+1$ players can be taken for $M \cup \{j\}$. If some players do trade in a^* then take one of these players to be player j and include some of his trading partners in M.

LINIAL, N. (1994), "Games Computers Play: Game Theoretic Aspects of Computing", in R. J. Aumann and S. Hart (eds.) *Handbook of Game Theory with Economic Applications* (New York: North-Holland).
MASKIN, E. (1999), "Nash Implementation and Welfare Optimality", *Review of Economic Studies*, **66**, 23–38.
MOORE, J. (1992), "Implementation, Contracts and Renegotiation in Environments with Complete Information", in J. J. Laffont (ed.) *Advances in Economic Theory Sixth World Congress* (Cambridge: Cambridge University Press).
SEGAL, I. (1999), "Contracting with Externalities", *Quarterly Journal of Economics*, **114**, 337–388.
SJOSTROM, T. (1993), "Implementation in Perfect Equilibria", *Social Choice and Welfare*, **10**, 97–106.